Introducing FoxPersonTracks: A benchmark for person re-identification from TV broadcast shows

Rémi Auguste, Pierre Tirilly, Jean Martinet

Introducing FoxPersonTracks: a Benchmark for Person Re-Identification from TV Broadcast Shows

Rémi Auguste, Pierre Tirilly, Jean Martinet
CRIStAL – Lille 1 University – Villeneuve d'Ascq, France
Email: {remi.auguste,pierre.tirilly,jean.martinet}@univ-lille1.fr

***Abstract*—This paper introduces a novel person track dataset dedicated to person re-identification. The dataset is built from a set of real life TV shows broadcasted from BFMTV and LCP TV french channels, provided during REPERE challenge. It contains a total 4,604 *persontracks* (short video sequences featuring an individual with no background) from 266 persons. The dataset has been built from the REPERE dataset by following several automated processing and manual selection/filtering steps. It is meant to serve as a benchmark in person re-identification from images/videos. The dataset also provides re-identifications results using space-time histograms as a baseline, together with an evaluation tool in order to ease the comparison to other re-identification methods.**

I. Introduction

Person re-identification from video is a fundamental task consisting in identifying an individual across shots of a single video or across various videos. It is an essential component of multi-camera tracking and has applications in automated videosurveillance, robotics and human computer-interaction, multimedia retrieval, and forensics.

Re-identifying individuals from real-life videos requires algorithms that are robust to changes in light, viewpoint, scale, and target appearance (due to gestures and, in some use cases, clothes, hair style, etc.). Developing and evaluating such algorithms requires benchmarks that offer data and evaluation protocols that encompass all the features of real-life data and all application domains. Current benchmarks for person re-identification from videos are limited in their range of applications (videosurveillance, mostly), size and settings. Additionally, recent work has shown the usefulness of re-identification for multimedia data understanding [1], calling for more research in this direction, and, consequently, dedicated benchmarks.

This paper presents a new dataset dedicated to the training and evaluation of methods for person re-identification in TV broadcast shows. This dataset is built from a set of real-life TV shows broadcasted from BFMTV and LCP TV french channels and provided during the REPERE challenge [2], and is available for download on ELDA website (URL $elda.fr$).

The paper is organized as follows : Section II describes existing re-identification datasets. Section III introduces the proposed dataset by describing its detailed content. Section IV presents experimental results on re-identification using space-time histograms, that are meant to serve as a baseline for this dataset. Section V concludes our work.

II. Related work

To our knowledge, all the existing datasets dedicated to re-identification bench-marking are based on videosurveillance data [3], with the exception of the Rope dataset [4], based on the movie *Rope* by A. Hitchock.

Most datasets [5] do not provide actual video data but only frames (eventually cropped) extracted from videos. In these datasets, the re-identification problem is set as follows: given a set of images, find the images that depict a single given individual. Based on this statement, the datasets only differ by their acquisition conditions (illumination, occlusions, viewpoint, etc. – see [3] for details) and their statistics (provided in Table I): number of individuals to be re-identified, number of images per individual, presence or not of distractor images (images that do not contain re-identifiable individuals). Although it focuses on the problem at hand, this definition of the re-identification problem:

- does not allow to benchmark re-identification algorithms based on motion features;
- offers a limited range of acquisition conditions: illumination, occlusions, poses, etc.;
- prevents the evaluation of the impact of potential pre-processing steps (e.g. body/face detection, background subtraction) on the re-identification performance.

Besides, as shown in Table I, they are of limited scale, either in terms of persons to be re-identified, or number of samples per person: none offers both a large amount of persons and a large amount of test cases per person.

Dataset	# Persons	# Images/pers.	Distractors
ViPER [5]	632	2	None
ETHZ [6]	146	59 (average)	None
i-LIDS for Re-ID [7]	119	4 (average)	None
CAVIAR4REID [8]	72	17 (average)	None
iLIDS-MA [9]	40	46	None
iLIDS-AA [9]	100	108 (average)	None
V-47 [10]	47	16	None
QMUL [11]	250	2	775
Rope [4]	8	55 (average)	None

TABLE I. Properties of existing re-identification frame-based benchmarks.

The datasets that are the closest to our work [12], [13] are both dedicated to face recognition in videos. They are based on different episodes from the TV series *Buffy the*

Vampire Slayer. These datasets provide face tracks, i.e. video sequences featuring a single face, labeled by the name of the corresponding character. Their statistics are detailed in Table II: although they provide real-world data, the number of persons to be (re-)identified is small, allowing to estimate the performance of the algorithms on a limited scale only.

Finally, one should also mention the *Hannah* dataset [14], dedicated to face tracking. Like the two *Buffy* datasets, it provides facetracks annotated by persons' names. Although it includes a fairly high amount of facetracks and persons (see Table II), the fact that it is based on a single movie makes the facetracks poorly representative of the complexity of its use case.

Dataset	# Persons	# Face tracks
Buffy (INRIA) [12]	8 (+1)	639
Buffy (Oxford) [13]	12 (+2)	1,526
Hannah [14]	238 (+16)	2,002

TABLE II. PROPERTIES OF EXISTING BENCHMARKS FOR FACE RECOGNITION IN VIDEOS. NOTE THAT "(+N)" DENOTES ADDITIONAL NON-CHARACTER-SPECIFIC CLASSES ("OTHERS", "CROWDS", "KIDS", "UNIDENTIFIED", AND "FALSE POSITIVES").

As compared to existing benchmarks, our dataset:

- covers an original yet significant use case of re-identification: TV broadcast shows;
- provides a large number of tracks (over 4,600) and a large number of individuals to be re-identified (266);
- provides full video shots rather than individual images, allowing to evaluate motion-based methods in addition to visual-based ones;
- is based on real-world data: excerpts from actual French TV broadcast shows.

III. DATASET DESCRIPTION

The main objective of this dataset is to evaluate re-identification algorithms, i.e. algorithms that aim at grouping shots that feature the same person (without explicitly naming this person). Baseline for face and upper body detection and segmentation provide a common ground to compare the absolute performance of various algorithms. Related tasks that can be evaluated using this dataset includes: person identification from video, face detection from video, relative impact of external components on the final re-identification performance (face detection, upper body detection, face/upper body segmentation).

A. Dataset contents

The proposed dataset is composed of:

- 4,604 video sequences from various TV shows, each featuring one of 266 different persons; such a sequence is called *persontrack* in the remainder of the paper;
- ground truth data providing the full name of the person for each shot;
- ground truth data providing the face position for a subset of 2081 keyframes extracted from the shots;
- baseline face detections for all frames of the shots;
- baseline detections of faces;
- baseline background subtraction data based on the detected face;
- evaluation software to compute the metrics presented in Section IV.

All data referred to as *ground truth* have been obtained manually. Data referred to as *baseline* as been obtained automatically.

In the remainder of this section, the source of the data and the processes used to produce the ground truth and baseline data are described, then statistics about the dataset are provided.

B. Data source

The original data used in this work was distributed as part of the REPERE challenge [2], which aimed at providing a benchmark for person identification from broadcast TV shows. During the challenge, 299 videos were released, that covered 9 TV shows originally broadcasted by the French TV channels LCP and BFMTV. This data came together with manually obtained ground truth for speech recognition and speaker recognition over the whole data, and OCR and face detection over a number of keyframes.

The dataset described here has been produced based on 134 videos from the REPERE dataset (videos available during phases 0 and 1 of the challenge) and their associated ground truth.

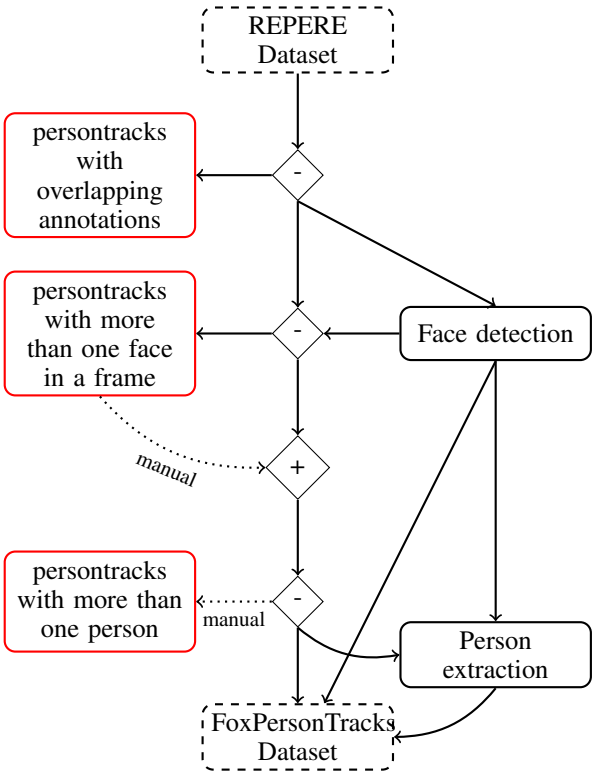

Fig. 1. Steps for building the presented dataset from the REPERE dataset.

C. Persontrack extraction and filtering

The process used to generate the persontracks from the original videos is illustrated in Figure 1. First, the REPERE ground truth is used to extract sequences of consecutive frames annotated with names. The ground truth directly provides the first and last frames of such sequences together with the name of the person appearing in it. Two or more of these sequences may overlap, meaning that more than one person is present in them. Also, some persons present in the sequence may not be annotated (e.g. persons in the audience) according to REPERE's annotation guidelines. The remaining steps of this process aim at removing these sequences to provide "pure" persontracks containing only a single person.

Overlapping sequences are then removed from the dataset, based on the ground truth information. After this second step, the only sequences that may contain more than one person are the ones that were not fully annotated. To deal with such sequences, faces are counted using a face detector based on the Viola and Jones algorithm [15] and a post-processing step that filters out false positives based on color. Every frame in which more than one face was detected is manually inspected: if the frame actually contains more than one face, the whole corresponding sequence is discarded. After this third step, the only remaining sequences that may contain more than one person are those on which the ground truth was incomplete and the detector failed. All these sequences are manually inspected and every sequence containing more than one person is discarded. This last filtering step ensures the accuracy of the dataset. At the end of the filtering, we obtain 4,604 persontracks of 266 different persons.

D. Baseline data generation

The face detections provided in our dataset are obtained using the Viola and Jones detector [15], and an additional post-processing step that computes the ratio of skin color in the detection window to filter out false positives and to resize the detection windows so that they contain as little background pixel as possible. Figure 2 shows an example of a detection window before and after this post-processing step.

After faces have been detected, the upper bodies are extracted based on a simple extension of the detection window. Then, the Grabcut algorithm is used to subtracts the background from the resulting windows (containing both the face and upper body). Figure 3 shows an example of these person-centered, background-free facetracks that are provided as baseline data in our dataset.

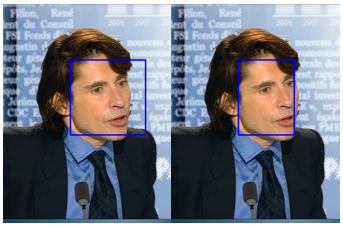

Fig. 2. Example of face detection before and after our optimization based on the skin color proportion.

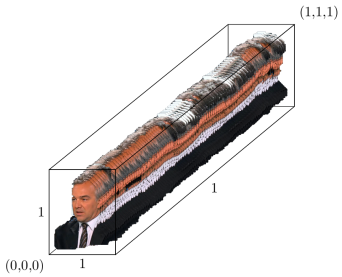

Fig. 3. Example of extracted persontrack from a BFMTV TV show.

E. Dataset statistics

In total, the corpus is composed of 266 different named persons seen in 134 videos. In average, there are 5 individuals per video (min 1, max 12), each individual appears in 1.2 TV shows among 9 (min 1, max 4), and each identity appears in 1.7 videos (min 1, max 15).

Figure 4 shows the occurrence distribution in a log-scale. Since the original videos are broadcast TV shows, the journalists (anchorpersons) appear more frequently than any other person. Some journalists appear in more than 50 shows, while many other persons appear only once.

Figure 5 shows the persontrack length distribution in the dataset. It shows that most persontracks are short in length, with an average of 55 frames. Note that as the peak shows, one third of the persontracks has a length between 31 and 38 frames.

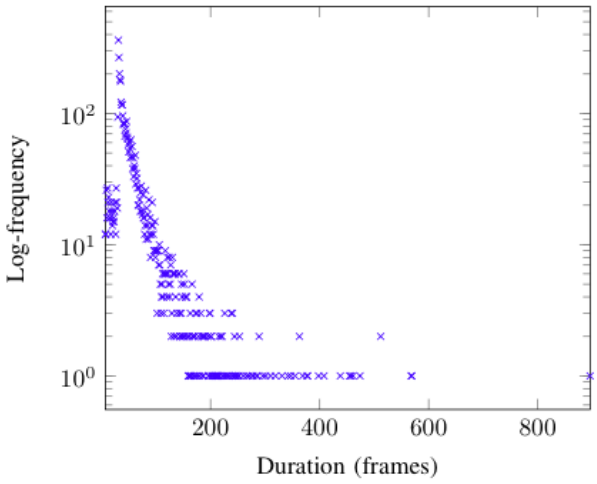

Fig. 4. Number of occurrences per identity distribution in the dataset ranging from 1 to 278 (x-axis) with an average of 17. The y-axis is logarithmic.

IV. Baseline results

The different steps leading towards our dataset have been evaluated. The face detection and face extraction algorithm were evaluated during the REPERE challenge using the official metrics[16]. We evaluated the results of our re-identification approach ourselves [17]. The first step of our re-identification approach is to build a space-time histogram (RGB, 1,500 bins) for each persontrack and then calculate a similarity matrix per

Fig. 6. Screenshots of representative persontracks featured in our dataset.

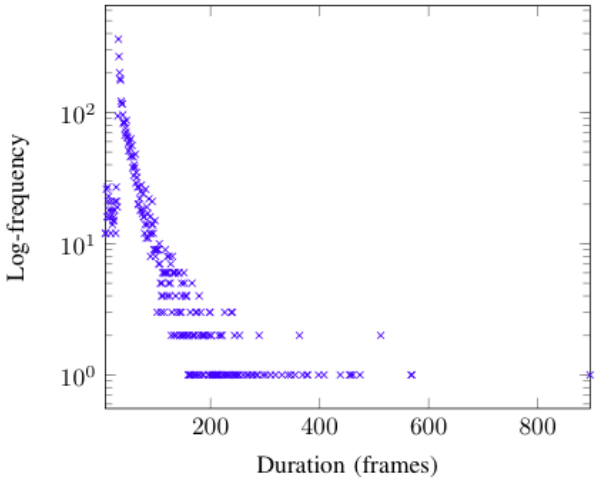

Fig. 5. Duration in frames distribution of our datasets ranging from 7 frames to 896 frames (x-axis) with an average of 55 frames. The y-axis is logarithmic.

show. The similarity measure for space-time histograms is a combination of the χ^2 and the mahalanobis distance (used as a similarity). A precision at n (P@N) [18] is then computed for each persontrack and the weighted average $\bar{P}$ is used as an evaluation measure:

$$\bar{P} = \frac{\sum_{i=1}^{|M|} P_i n_i}{\sum_{i=1}^{|M|} n_i} \quad (1)$$

where M is the similarity matrix, i are the persontracks, n_i the number of persontracks bearing the same identity than i and P_i is the precision at n_i.

Method	Precision	size of evaluation	# of correct
Face detection	0.44	2,081	923
Face extraction	0.54	923	498
Re-identification	0.898	4,604	4,071

TABLE III. COMPARISON OF THE PRECISION OF THE RE-IDENTIFICATION ALGORITHM.

V. CONCLUSION

We have introduced a new re-identification dataset named FoxPersonTracks. This dataset contains 4,604 persontracks showing 266 individuals, and compared to other existing datasets, it provides a large number of persontracks and a large number of individuals, with a wide range of acquisition conditions from several TV shows; also, it provides full video shots rather than individual images, allowing to use the motion. Finally, it provides an evaluation tool making it easy to compare algorithms in a unified way according to the presented metrics. We believe that this dataset is useful as a benchmark for person re-identification from images/videos.

REFERENCES

[1] B. Favre, G. Damnati, F. Bechet, M. Bendris, D. Charlet, R. Auguste, S. Ayache, B. Bigot, A. Delteil, R. Dufour, C. Fredouille, G. Linares, J. Martinet, G. Senay, and P. Tirilly, "Percoli: A person identification system for the 2013 repere challenge," in *First Workshop on Speech, Language and Audio for Multimedia (SLAM)*, 2013.

[2] O. Galibert and J. Kahn, "The first official repere evaluation," in *First Workshop on Speech, Language and Audio for Multimedia (SLAM)*, 2013.

[3] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, 2014.

[4] V. Gandhi and R. Ronfard, "Detecting and naming actors in movies using generative appearance models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3706–3713.

[5] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *IEEE Intl workshop on performance evaluation of tracking and surveillance*, 2007.

[6] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *Brazilian Symposium on Computer Graphics and Image Processing*, 2009, pp. 322–329.

[7] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proceedings of the British Machine Vision Conference*, 2009.

[8] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proceedings of the British Machine Vision Conference*, 2011.

[9] S. Bak, E. Corvee, F. Bremond, , and M. Thonnat, "Boosted human re-identification using riemannian manifolds," *Image and Vision Computing, Special Issue on Manifolds for Computer Vision*, vol. 30, no. 6-7, pp. 443–452, 2012.

[10] S. Wang, M. Lewandowski, J. Annesley, and J. Orwell, "Re-identification of pedestrians with variable occlusion and scale," in *International Workshop on Visual Surveillance (In Conjunction with ICCV)*, 2011, pp. 1876–1882.

[11] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1988–1995.

[12] R. G. Cinbis, J. Verbeek, and C. Schmid, "Unsupervised metric learning for face identification in tv video," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1559–1566.

[13] M. Everingham, J. Sivic, and A. Zisserman, ""Hello! My name is... Buffy" – automatic naming of characters in TV video," in *Proceedings of the British Machine Vision Conference*, 2006.

[14] A. Ozerov, J.-R. Vigouroux, L. Chevallier, and P. Perez, "On evaluating face tracks in movies," *IEEE ICIP*, pp. 3003–3007, 2013.

[15] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision (IJCV)*, vol. 57, no. 2, pp. 137–154, 2002.

[16] J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, and P. Joly, "A presentation of the repere challenge," in *CBMI*, 2012, pp. 1–6.

[17] R. Auguste, J. Martinet, and P. Tirilly, "Space-time histograms and their application to person re-identification in tv shows," in *ACM ICMR*, 2015, in press.

[18] V. Raghavan, P. Bollmann, and G. S. Jung, "A critical investigation of recall and precision as measures of retrieval system performance," *ACM Transactions on Information Systems (TOIS)*, vol. 7, no. 3, pp. 205–229, 1989.